Protein Engineering vol.8 no.6 pp.535–542, 1995

The predicted secondary structure of the G-type glutamine amidotransferase is compatible with TIM-barrel topology

Thomas Niermann[1,2] and Kasper Kirschner[1,3]

[1]Department of Biophysical Chemistry, Biozentrum der Universität Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

[2]Present address: Ciba-Geigy AG, Biotechnologie, CH 4002 Basel, Switzerland

[3]To whom correspondence should be addressed

Glutamine amidotransferase (GAT) subunits or domains catalyze an important partial reaction in many complex biosynthetic reactions. The structure of one member of the F-type GATs is known, but the structure of the unrelated G-type is still unknown. Because many protein sequences are available for anthranilate synthase component II (product of the *trp*G gene), we have predicted its average secondary structure by a joint prediction method [Niermann and Kirschner (1991a) *Protein Engng*, 4, 359–370]. The predicted eight β-strands and seven α-helices follow an 8-fold cyclic repetition of a β-strand–loop–α-helix–loop module with helix α_7 missing. This pattern of secondary structure suggests that the G-type GAT domain has an 8-fold βα-barrel topology, as found first in triose phosphate isomerase (TIM-barrel). This model is supported by the location of known catalytically essential residues in loops between β-strands and α-helices. Evidence from published sequencing and mutational studies on selected members of the GAT superfamily (carbamoyl phosphate, imidazoleglycerol phosphate, GMP and CTP synthases) support both the secondary structure prediction and the TIM-barrel topology.

Key words: glutamine amidotransferase/prediction of secondary structure/sequence alignments/TIM-barrel proteins/tryptophan biosynthesis

Introduction

More than a dozen biosynthetic enzymes possess both an ammonia (NH_3)-dependent synthase and a glutamine amidotransferase (GAT) activity. A recent review is given by Zalkin (1993). In general, the synthase incorporates either NH_3 or preferentially the amino group released from glutamine into a variety of intermediates in the biosynthesis pathways of amino acids, nucleotides and coenzymes. The two different active sites of these complex enzymes are provided for by different domains (or subunits), and are thought to interact intimately in terms of both structure and function. A striking consequence of this interdomain interaction is the failure of 'nascent NH_3', which is released from glutamine at the GAT active site, to equilibrate with the bulk solvent.

There are two apparently unrelated families of GAT domains, designated F- and G-type, for historical reasons. The structure of the first complex enzyme containing an F-type GAT domain has been determined recently for glutamine 5'-phosphoribosyl-1-pyrophosphate amidotransferase, the product of the *pur*F gene from *Bacillus subtilis* (Smith *et al.*, 1994). The F-type GAT domain consists of two antiparallel β-sheets positioned face to face, and covered by α-helices on both outer surfaces.

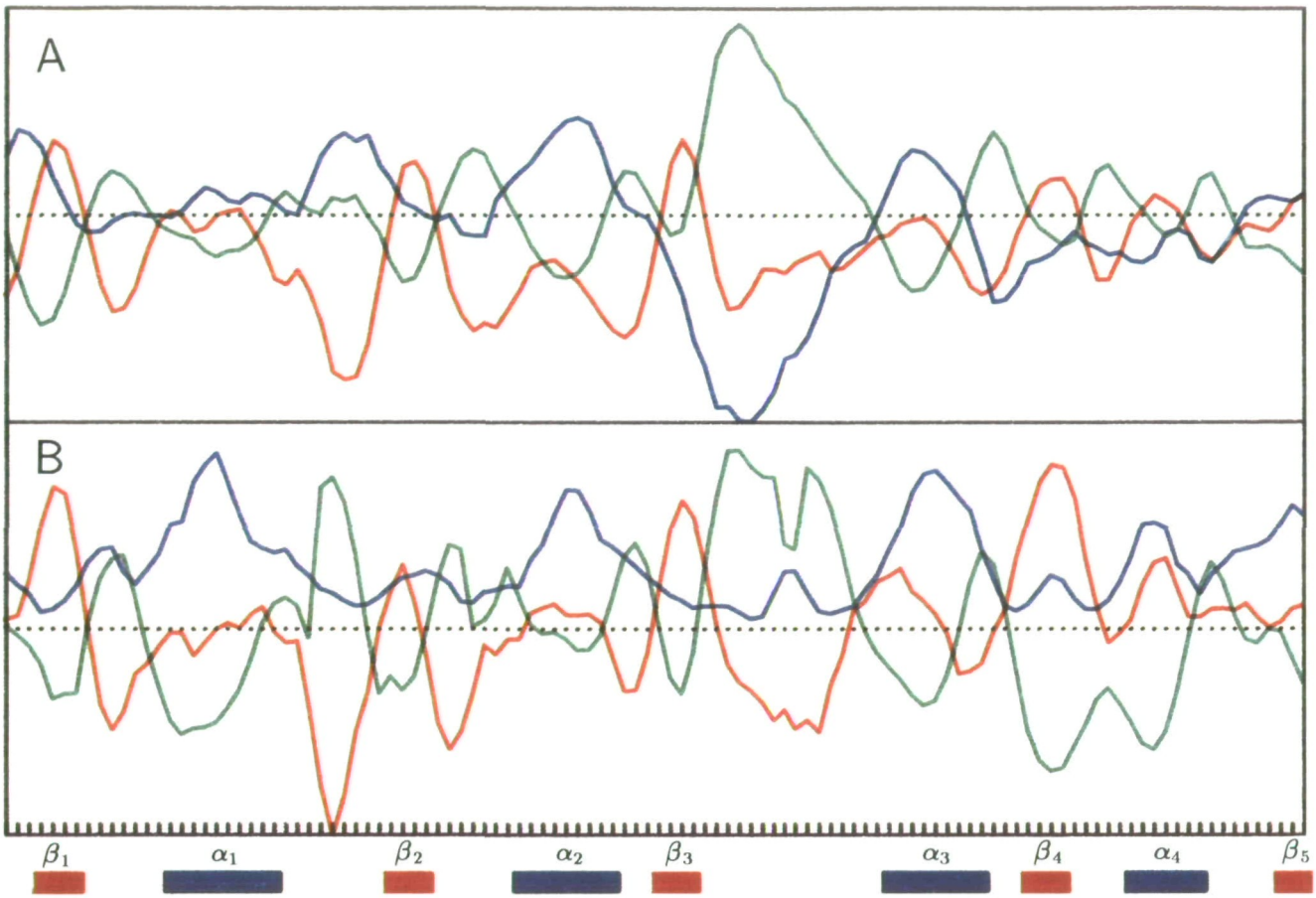
In contrast, the structure of the G-type GAT domain is still unknown. However, several predictions have tentatively assigned an entirely different protein fold, namely that of the familiar 8-fold parallel βα (or TIM)-barrel (Farber, 1993), to these domains (Wilmanns and Eisenberg, 1993; T.Niermann and K.Kirschner, personal communication to Wilmanns and Eisenberg, 1993). A preliminary report on the X-ray crystallographic structure analysis of GMP synthetase from *Escherichia coli* that contains a G-type GAT domain has been published recently (Tesmer *et al.*, 1994). The anticipated first determination of the structure of a G-type GAT domain provides for a test of the combination of methods suitable for the improved prediction of the secondary structure of α/β proteins (Niermann and Kirschner, 1991b), particularly that of TIM-barrel proteins (Crawford *et al.*, 1987; Niermann and Kirschner, 1991a).

Here we report the results of applying the methods to 26 sequences of the G-type GAT domains of both anthranilate synthase and the closely related aminodeoxychorismate synthase. The predicted average secondary structure consists of a sequence of eight βα modules, with helix α_7 predicted as a loop. The cysteine, histidine and glutamate residues known to be catalytically essential are located in loops following the C-termini of predicted strands β_4 and β_8 that would be spatially adjacent in a TIM-barrel fold. Data from aligned sequences of the superfamily of synthetases with G-type GAT domains narrow down the number of invariant residues at the carboxyl ends of β-strands still further, and support the TIM-barrel fold.

Materials and methods

Alignment of amino acid sequences

We consider here 22 *trp*G sequences (*Sma*, *Serratia marcescens*; *Eco*, *Escherichia coli*; *Lbi*, *Leptospira biflexa*; *Aca*, *Acinetobacter calcoaceticus*; *Pa1*, *Pa2*, *Pseudomonas aeruginosa* (two genes; Essar *et al.*, 1990); *Sty*, *Salmonella typhimurium*; *Bla*, *Brevibacterium lactofermentum*; *Tth*, *Thermus thermophilus*; *Ncr*, *Neurospora crassa*; *Sce*, *Saccharomyces cerevisiae*; *And*, *Aspergillus nidulans*; *Ang*, *Aspergillus niger*; *Pch*, *Penicillium chrysogenum*; *Sdy*, *Shigella dysenteria*; *Vpa*, *Vibrio parahaemolyticus*; *Bsu*, *Bacillus subtilis*; *Ppu*, *Pseudomonas putida*; *Rme*, *Rhizobium meliloti*; *Pbl*, *Phycomyces blakesleeanus*; *Pcp*, *Phanerochaete chrysosporium*; *Hpo*, *Hansenula polymorpha*; and four *pab*A sequences [*E.coli*, *S.typhimurium*, *Kae* (*Klebsiella aerogenes*) and *S.marcescens*]. All sequences were obtained from the MIPS data bank (Mewes, 1990). The sequences were first aligned in pairs, using FASTA (Pearson and Lipman, 1988; GCG Sequence Analysis Software Package, 1991). The computer programs PROFILE and PROFILEGAP (Gribskov *et al.*, 1987; GCG Sequence Analysis Software Package, 1991) were used for multisequence alignment. Minor rearrangements in the sequence alignments were greatly aided by the clustering of both identical residues and




```
                  10        20        30        40        50        60        70        80        90       100       110
                   |         |         |         |         |         |         |         |         |         |         |
cons.       DN DSFt N         g          RN                    SPGP  P     G          g  P  G C G Q      Gg
Sma MADILLLDNVDSFTYNLVDQLRASGHQ......VVIYRNQIGAEVIIERLQHMEQPVLMLSPGPGTP...SEAGCMPELLQRLRGQLPIIGICLGHQAIVEAYGGQVGQAGEI
Eco MADILLLDNIDSFTYNLADQLRSNGHN......VVIYRNHIPAQTLIERLATMSNPVLMLSPGPGVP...SEAGCMPELLTRLRGKLPIIGICLGHQAIVEAYGGYVGQAGEI
Lbi .MKVLILDNYDSFTFNLYQIVGEILEEREKPFQLDVIRND...EKPFEWIKSANYDKIIISPGPGHPADPAYFGVSADILKELGKTPPVLGICLGMQGMATVFGGEVVRANIA
Aca ..MLLMIDNYDSFTYNIVQYFGELNQD......VKVVRND...QVTLEDIERWQPKYLVVGPGPCSP...TEAGISIPAIHHFAGRIPLLGVCLGHQAIGQAFGGNIIRAKTV
Sty MADILLLDNIDSFTWNLADQLRTNGHN......VVIYRNHIPAQTLIDRLATMKNPVLMLSPGPGVP...SEAGCMPELLTRLRGKLPIIGICLGHQAIVEAYGGYVGQAGEI
Bla MTHVVLIDNHDSFVYNLVDAFAVAGYK......CTVFRNT....VPVETILAANPDLICLSPGPGYP...ADAGNMMALIERTLGQIPLLGICLGYQALIEYHGGKVEPCG.P
Tth VMRVLVVDNYDSFTYNLVQYLGELGAE......PIVWRND...RFRLEEVEALDPDRILISPGPCTP...FEAGLSVPLVQRYAPRYPILGVCLGHQAIGAAFGGKVVPAPVL
Sdy MADILLLDNIDSFTYNLADQLRSNGHN......VVIYRNHIPAQTLIERLATMSNPVLMLSPGPGVP...SEAGCMPELLTRLSGKLPIIGICLGHQAIVEAYGGYVGQAGEI
Vpa MANIVFIDNFDSFTYNLVDQFRSLGHS......VKIYRNHIPAETIEQAINELENPVVLLSPGPGAP...SEAGSMPELIQRMKGKVPMIGICLGHQAIVEAYGGTVAGAGEI
Bsu ..MILMIDNYDSFTYNLVQYLGELGEE......LVVKRND...SITIDEIEELSPDFLMISPGPCSP...DEAGISLEAIKHFAGKIPIFGVCLGHQSIAQVFGGDVVRAERL
Sce NKHVVLIDNYDSFTWNVYEYLCQEGAK......VSVYRND...AITVPEIAALNPDTLLISPGPGHP..KTDSGISRDCIRYFTGKIPVFGICMGQQCMFDVFGGEVAYAGEI
Hpo SKNVVMIDNYDSFTWNLYEYLCQEGAN......VEVFRND...QITIPEIEQLKPDVVVISPGPGHP..RTDSGISRDVISHFKGKIPVFGVCMGQQCIFEEFGGDVEYAGEI
And ASNVILIDNYDSFTWNVYQYLVLEGAT......VTVIRND...EISLEELIAKKPTQLVVSPGPGHP..KSDAGISNAAIQYFAGKIPIFGVCMGQQCIHHSFGGKVDVTGEI
Ang ASNVILIDNYDSFTWNVYQYLVLEGAT......VNVFRND...QITLEELIAKKPTQLVISPGPGHP..ETDAGISSAAIQYFSGKIPIFGVCMGQQCIITCFGGKVDVTGEI
Pch ASNVILIDNYDSFTWNIYQYLVLEGAT......VTVYRND...EVTVEDLVAKKPTQLVISPGPGHP..DTDAGISNAVIKHFSGKVPIFGVCMGQQCMITSFGGKVDVTGEI
Ncr ASNLILIDNYDSFTWNVYQYLVLEGAK......VTVFRND...QITIDELIAKNPTQLVISPGPGHP..GTDSGISRDAIRHFAGKIPIFGVCMGQQCIFDVYGGDVCFAGEI
Pbl .MATLLIDNYDSFTYNVYQYLCSQGAD......VVVYRND...KITVDEIVKLNPVNIVISPGPGHP..SHDAGVSRDVISYFAGKLPILGICMGEQCIFEVFGGTVSYAGDI
Pcp PIDILMIDNFDSFTWNLYQSLCLLGAD......VTVIRND...AIPRSAIPQLRIKRLIVSPGPGHP..QTDSGISREAIKYFAGKVPIMGVCMGLECVVDVFGGQIAYAGEI
Rme GVSILLVDHEDSFVHTLANYFRQTGAS......VTTVRTP....VAEEIFDRVKPDLVVLSPGPGTP...KDFDCKATIKKARARDLPIFGVCLGLQALAEAYGGDLRQLAIP
Pa1 ..MLLMIDNYDSFTYNLVQYFGELKAE......VKVVRND...ELSVEQIEALAPERIVLSPGPCTP...NEAGVSLAVIERFAGKLPLLGVCLGHQSIGQAFGGEVVRARQV
Pa2 ..MLLMIDNYDSFTYNVVQYLGELGAE......VKVIRND...EMTIAQIEALNPERIVVSPGPCTP...SEAGVSIEAILHFAGKLPILGVCLGHQSIGQAFGGDVVRARQV
Ppu .MRITLLDNFDSFTYNLVEQFCLLGAE......VRVMRNDTPLPTIQAALLADGCELLVLSPGPGRP...EDAGCMLELLAWARGRLPVLGVCLGHQALALAAGGAVGEARKP
Eco ..MILLIDNYDSFTWNLYQYFCELGAD......VLVKRND...ALTLADIDALKPQKIVISPGPCTP...DEAGISLDVIRHYAGRLPILGVCLGHQAMAQAFGGKVVRAAKV
Sty ..MILLIDNYDSFTWNLYQYFCELGAE......VQVRRND...ALTLAHIDALNPQKIVISPGPCTP...NDAGISLAVIRHYAGRIPMLGVCLGHQAMAQAFGASVVRAAKV
Kae ..MILLIDNYDSFTWNLYQYFCELGAE......VLVRRND...ELTLADIISLAPAKIVISPGPCTP...DESGISLAAIRHFSGQTPILGVCLGHQAIAQVFGAAIVRAAKV
Sma ..MLLLIDNYDSFTYNLYQYFCELGAE......VVVKRND...ELQLTDIERLAPQHLVISPGPCTP...NEAGISVAAIRHFAGKLPILGVCLGHQALGQAFGAEVVRARAV

var.122                  111  2       4 2      121212 3113 34        2   33   1 311 3111  1         1    221    3 22 111
    85696841291112824668548758011111151461272229373597774390369621113011163642249446963458416913131027620161294635390
```

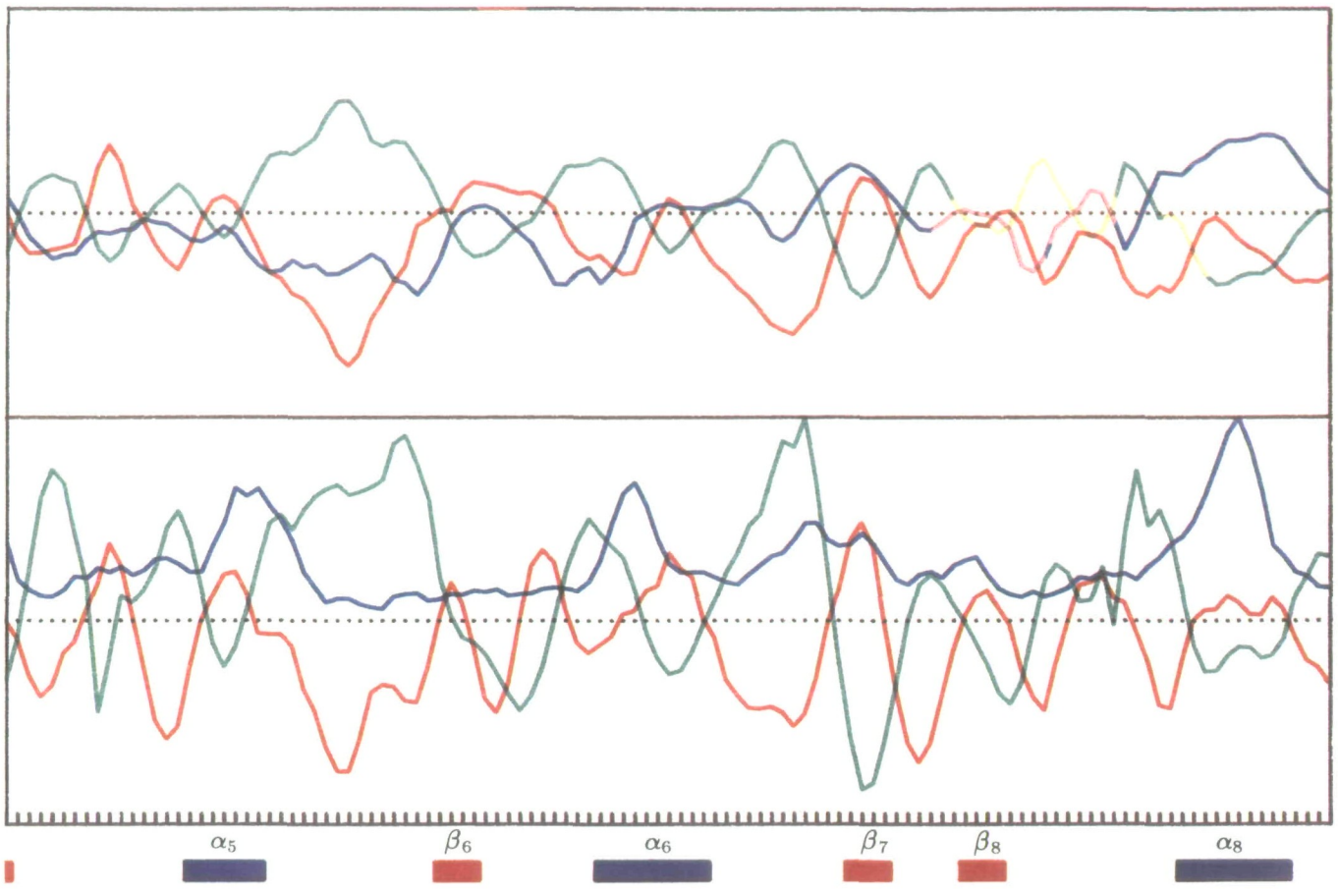
gaps throughout the sequences, which provided the necessary internal register for confident alignment.

Averaging of secondary structure propensities

The secondary structure 'GOR' prediction algorithm [Garnier *et al.* (1978), with decision constants as suggested by Gibrat *et al.* (1987)] was first used to calculate the propensity of each residue for the α-helix, β-strand and coil state along each continuous amino acid sequence. Then the propensities corresponding to each residue were located to that residue's position in the alignment. The average state propensities at each position were obtained by summing up and dividing the sum by the number of amino acids in that column (Garnier *et al.*, 1978). The 'state profiles' were finally smoothed with a three-residue span. The state with the highest average propensity defines the predicted secondary structure at that position.

Averaging of residue properties

The chain flexibility profiles of individual sequences were calculated according to Karplus and Schulz (1985). Hydropathy profiles were calculated using the scale of Kyte and Doolittle (1982), over a five-residue span setting, and were finally




```
      120       130       140       150       160       170       180       190       200       210       220       230
      |         |         |         |         |         |         |         |         |         |         |         |
 HGk s      h                        RYHSL     P                   M   h     g QFHPES      G
LHGKASAI....AHDGEGMFAGMANP.............LPVARYHSLVGS..NIPADLTVNARSG.......EMVMAVRDDRRRVCGFQFHPESILT...THGARLLEQTLAWALAK
LHGKASSI....EHDGQAMFAGLTNP.............LPVARYHSLVGS..NIPAGLTINAHFN.......GMVMAVRHDADRVCGFQFHPESILT...TQGARLLEQTLAWAQHK
MHGKLSPI....EHDGKGVFSGLTQG.............IEIMRYHSLVAKEISLPNDLEITARVSAGEGK..GEIMGLRHKSLKIEGVQFHPESFGS...EEGKCLLRNFINS....
MHGRLSDM....YHTDKGIFSNLPSP.............FSATRYHSLVIEQESLPECLEVTCWTNQNDGS.IEEIMGVKHKTLPVEGVQFHPESILS...QHGHQIFKNFLEIYA..
LHGKASSI....EHDGQAMFAGLANP.............LPVARYHSLVGS..NVPAGLTINAHFN.......GMVMAVRHDADRVCGFQFHPESILT...TQGARLLEQTLAWAQQK
VHGTTDNMILTDAGVQSPVFAGLATDVEPDHPEVPGRKVPIGRYHSLGCV..VAPDGIESLGTCSSEI...GDVIMAARTTDGKAIGLQFHPESVLS...PTGPIILSRCVEQLLAN
MHGKVSPI....HHDGTGVFRGLDSP.............FPATRYHSLAVV..EVPEALVVNAWAEEA...GGRTVMGFRHRDYPTHGVQFHPESYLT...EAGKLILKNFLEDPWTR
LHGKASSI....EHDGQAMFAGLTNP.............LPVARYHSLVGS..NIPAGLTINAHFN.......GMVMAVRHDADRICGFQFHPESILT...TQGARLLEQTLAWAQRK
IHGKVSMM....EHQDHAIYQNLPSP.............LAIARYHSLVAT..KVPDSLTITAEVD.......NLVMSVVHEQDKVCGFQFHPESIMT...TYGATLLGNAIEWALEK
MHGKTSDI....EHDGKTIFEGLKNP.............LVATRYHSLIVKPETLPSCFTVTAQTKE......GEIMAIRHNDLPIEGVQFHPESIMT...SFGKEMLRNFIETYRKE
VHGKTSPI....SHDNCGIFRNVPQG.............IAVTRYHSLAGTESSLPSCLKVTASTE......NGIIMGVRHKKYTVEGVQFHPESILT...EEGHLMIRNILNVSGGT
VHGKTSTV....KHDNKGMFKNVPQD.............VAVTRYHSLAGTLKSLPDCLEITARTDN......GIIMGVRHKKYTIEGVQFHPESILT...EEGHLMIQNILNVSGGY
LHGKTSVL....KHDGRGAYEGLPPS.............VIITRYHSLAGTHSTIPECLEVSSFAQLGEDADKTVIMGVRHKQFAVEGVQFHPESILT...EHGQTMFRNFLKLTAGT
LHGKTSPL....KHDGKGAYEGLPGS.............LAVTRYHSLAGTHATIPDCLEVSSSVQLADDSNKDVIMGVRHKKLAVEGVQFHPESILT...EYGRIMFRNFLKLTAGT
LHGKTSEL....KHDSKGVYQGLPTS.............LEVTRYHSLAGTHSTIPDCLEVTSRVELGDASGKNIIMGVRHKEFAVEGVQFHPESILT...QYGRKMFRNFLELTAGT
LHGKTSPL....RHDGKGAYAGLSQD.............LPVTRYHSLAGTHVTLPECLEVTSWIAKEDG.SKGVIMGVRHKEYTIEGVQFHPESILS...AEGRGMFRNFLHMQGGT
LHGKTSTI....KHDNRGLFKNVPQD.............NQVTRYHSLAGMPSTLPEVLEVTATTDD......GVIMGVRHKKYTVEGVQFHPESILC...EHGHTMISNFLSLRGGN
MHGKVSGI....RHDARGCFKDLPQG.............IQSTRYHSLSAGVKTLPDELAVTAVTEHE.....RVIMGIRHRKYTVEAVQYHPESILS...ESGDDLLRHFMKLKGGT
MHGKPSRI...RVLEPGIVFSGLGKE.............VTVGRYHSIFADPSNLPREFVITAESED......GTIMGIEHSKEPVAAVQFHPESIMTLGGDAGMRMIENVVAHLAKR
MHGKTSPI....HHKDLGVFAGLANP.............LTVTRYHSLVVKRESLPECLEVTAWTQHADGS.LDEIMGVRHKTLNVEGVQFHPESVLT...EQGHELLANFLRQQGGV
MHGKTSPV....HHRDLGVFTGLNNP.............LTVTRYHSLVVKRETLPDCLEVTAWTAHEDGS.VDEIMGLRHKTLNIEGVQFHPESILT...EQGHELFANFLKQTGGR
LHGKSTSL...RFDQRHPLFDGIAD...............LRVARYHSLVVS..RLPEGFDCLAESD.......GEIMAMADPRNRQLGLQFHPESILT...THGQRLLENALLWCGAL
MHGKTSPI....THNGEGVFRGLANP.............LTVTRYHSLVVEPDSLPACFDVTAWSET......REIMGIRHRQWDLEGVQFHPESILS...EQGHQLLANFLHR....
MHGKTSPV....THNGQGVFRGLPSP.............LTVTRYHSLIVDPATLPECFEITAWSET......QEIMGIRHREWDLEGVQFHPESILS...EQGHALLKNFLRR....
MHGKTSPV....SHTGQGVFLGLNNP.............LTVTRYHSLLIDPRTLPECFEVTARSEE......GEIMGIRHRVFDLEGVQFHPESILS...EQGHQLLANFLNR....
MHGKTSAI....RHLGVGVFRGLSDP.............LTVTRYHSLVLKADTLPDCFEVTAWSERDGVR..DEIMGIRHRALALEGVQFHPESVLS...EQGHQLLDNFLNR....

    1 2     4 114 1 3  22              3       113221  11 1   22271   1111   1  13331              12 24  2 1 344112
91130397111335651742345139111111111111183661111254049761734166593305776055631416399362924121111541117331186565055826 67
```

Fig. 1. Joint prediction of average secondary structure of the G-type GAT. Section of aligned sequences. Top row 'cons': upper case, invariant residues, allowing for one possible sequencing error; lower case, highly conserved residues, allowing for three differences. There was a total of 22 sequences of anthranilate synthase component II (*Sma* to *Ppu*; for definition of acronyms of microorganisms see Materials and methods) and four sequences of aminodeoxychorismate synthase (*Eco* to *Sma*). The single-letter amino acid codes have been used. Numbering begins with M1 of GAT from *S.marcescens*, and is continuous throughout the longest sequences across the gaps. Bottom row 'var': sequence variability index (see Materials and methods). Upper row, tens; lower row, ones. (**A**) Profiles of averaged secondary structure propensity. Color code: blue, α-helix; red, β-strand; green, coil. (**B**) Profiles of averaged property parameters. Color code: blue, amphipathic moment; red, hydropathy; green, chain flexibility. Joint prediction of secondary structural elements: blue bars (helices α_i) and red bars (strands β_i) separated by coil segments. See text for details.

smoothed with a three-residue span. The scale of Kyte and Doolittle (1982) (Cornette *et al.*, 1987) was also used for the calculation of the helical amphipathic moment at 100° (Eisenberg *et al.*, 1984), with a span setting of seven residues and a final smoothing with a three-residue span. The chosen span setting of seven residues for the amphipathic moment

profile is shorter than the value we originally preferred (Niermann and Kirschner, 1991a). The justification here is that the central amphipathic regions of some predicted helices seem to be rather short, therefore a longer span setting would yield less pronounced maxima. The average 'property profiles' of the aligned sequences were obtained as described above for the state profiles. Peaks of the property profiles were used in the joint prediction procedure (Niermann and Kirschner, 1991b) to enhance the cognate propensity profile, where appropriate.

The variability index '*var*' (Wu and Kabat, 1970) is a measure of sequence variability. $var = m/f$, where m is the number of different residues, and f ($0 < f < 1$) the fraction of the most frequent residue at that position. It follows that $1 < var < n^2$, where n is the number of aligned sequences.

Results and discussion

Information from aligned sequences

The bottom of Figure 1 presents 26 aligned sequences of G-type GAT domains. The first 22 sequences represent component II of anthranilate synthase (i.e. products of the *trp*G gene), and the last four sequences correspond to the closely related 4-amino-4-deoxychorismate synthase (*pab*A gene). The presented alignment spans the entire sequence of the monomeric GAT subunit from *S.marcescens* (Tso and Zalkin, 1980) to eliminate irrelevant extensions at the N- and C-termini. The sequences are numbered continuously across the gaps for later comparison with the sequences of the superfamily of GAT domains. Similar alignments, albeit with fewer sequences, have been published previously (Kaplan *et al.*, 1985; Essar *et al.*, 1990) as evidence for the evolutionary relationship between the products of the *trp*G and *pab*A genes. Here the emphasis is on the information content of aligned sequences of a G-type GAT domain regarding the probable occurrence of secondary structural elements along the sequence.

There are eight aligned gaps, indicating that the protein fold tolerates insertions and deletions at these sites. It is generally accepted that the residues that border the gaps reside in loops on the protein surface (Zvelebil *et al.*, 1987). This assignment is reinforced by the preferred occurrence of Pro, Gly and charged residues in the segments that span the gaps. The gaps divide the set of aligned sequences into nine blocks that probably contain the uninterrupted secondary structural elements of the protein core (Niermann and Kirschner, 1991a,b; Benner, 1992).

The row labeled 'cons' above the block of aligned sequences denotes the 36 out of 230 (16%) conserved residues. The eight clusters of conserved residues reported previously by Kaplan *et al.* (1985) and Essar *et al.* (1990) are preserved despite the larger number of sequences. However, the number of conserved residues has decreased, focusing on those that are important for folding and catalysis. The clusters are numbered as follows: (I) 8 DNxDSFt 14, (II) 38 RN 39, (III) 61 SPGxxP 67, (IV) 91 GxCxGxQ 97, (V) 104 Gg 106, (VI) 115 HGk 117, (VII) 156 RYHSL 160, (VIII) 202 QFHPES 207, where x is a variable residue position.

The two rows below the block of aligned sequences give the sequence variability at each position, as defined by Wu and Kabat (1970). These data confirm the earlier work of Crawford (1989), showing that regions of high and low variability alternate.

Prediction of secondary structure

Figure 1A displays the three profiles of averaged secondary structural propensities based on the GOR method (Gibrat *et al.*, 1987; red, β-strand; blue, α-helix; green, coil). We have shown previously (Niermann and Kirschner, 1991a,b) that averaging of propensities is preferable to consensus prediction, and that the accuracy of prediction increases with sequence variability. It is seen that the G-type GAT domain belongs to the α/β class of proteins (Levitt and Chothia, 1976) because β-strands tend to alternate with α-helices. Figure 1B displays the profiles of averaged residue properties that correlate with β-strands (red, hydrophobicity; Kyte and Doolittle, 1982), α-helices (blue, amphipathic moment; Eisenberg *et al.*, 1984) and coils (green, chain flexibility; Karplus and Schulz, 1985). In general, the cognate state propensity and property profiles vary synchronously across the entire alignment, and are therefore mutually supportive.

We have used the qualitative approach as described by Niermann and Kirschner (1991b) to obtain a joint prediction of average secondary structural elements along the protein sequence. The state propensity profiles represent the primary information. Where the amplitudes of the helix and strand propensities are weak and of identical sign and magnitude, the prediction is guided qualitatively by the stronger of the correlated property profiles. The joint prediction is presented below Figure 1B as a sequence of blue (α-helix) and red (β-strand) bars, with the intervening spaces assigned to surface loops.

Ignoring the profiles across the aligned gaps, the alternating patterns of hydrophobicity and chain flexibility in Figure 1B strongly support the prediction in Figure 1A of five β-strands (β_1, β_2, β_3, β_4 and β_7). Similarly, the maxima of amphipathic moment reinforce the prediction of five α-helices (α_1, α_2, α_3, α_6 and α_8). Strands β_6 and β_8 are so predicted because the maxima of strand propensity and hydropathy are more pronounced than those of helix propensity and amphipathic moment. Strand β_5 should be assigned to an α-helix by these criteria, but its β-strand character is justified by applying the template criterion as discussed later on. Similarly, the assignment of α_5 is based on the strong peak of amphipathic moment. In summary, up to this point only the assignments of β_5 and α_4 remain ambiguous; α_7 is missing, and there is the indication of an additional β-strand overlapping with the C-terminus of α_6.

Evidence in favor of a TIM-barrel fold

The pattern of strongly predicted β-strands and α-helices suggests that the G-type GAT domain has the TIM-barrel topology. This fold consists of a cyclic 8-fold repeat of the β-strand–loop–α-helix–loop module. The β-strands assemble to a central eight-stranded parallel β-barrel connected on the outside by α-helices. All TIM-barrels with a known enzymatic function have their active sites at the C-terminal face of the central β-barrel (Farber, 1993), with catalytic residues located in the C-terminus of several β-strands and in the loops that connect these β-strands to the next α-helix. The overall length of the polypeptide chain ($190 < n < 241$ residues), as well as the average lengths of the predicted β-strands (five ± one residues) and α-helices (10 ± two residues), are consistent with the TIM-barrel topology (Lesk *et al.*, 1989; Lasters *et al.*, 1990; Murzin *et al.*, 1994a,b).

Figure 2 depicts the predicted secondary structural elements of Figure 1 arranged into an idealized TIM-barrel. The triangles represent β-strands rising towards the observer. They are connected to the next α-helix (large circles labeled α_i) by strings of small circles representing residues of the connecting loops. The amino acids (single-letter code) contained in the

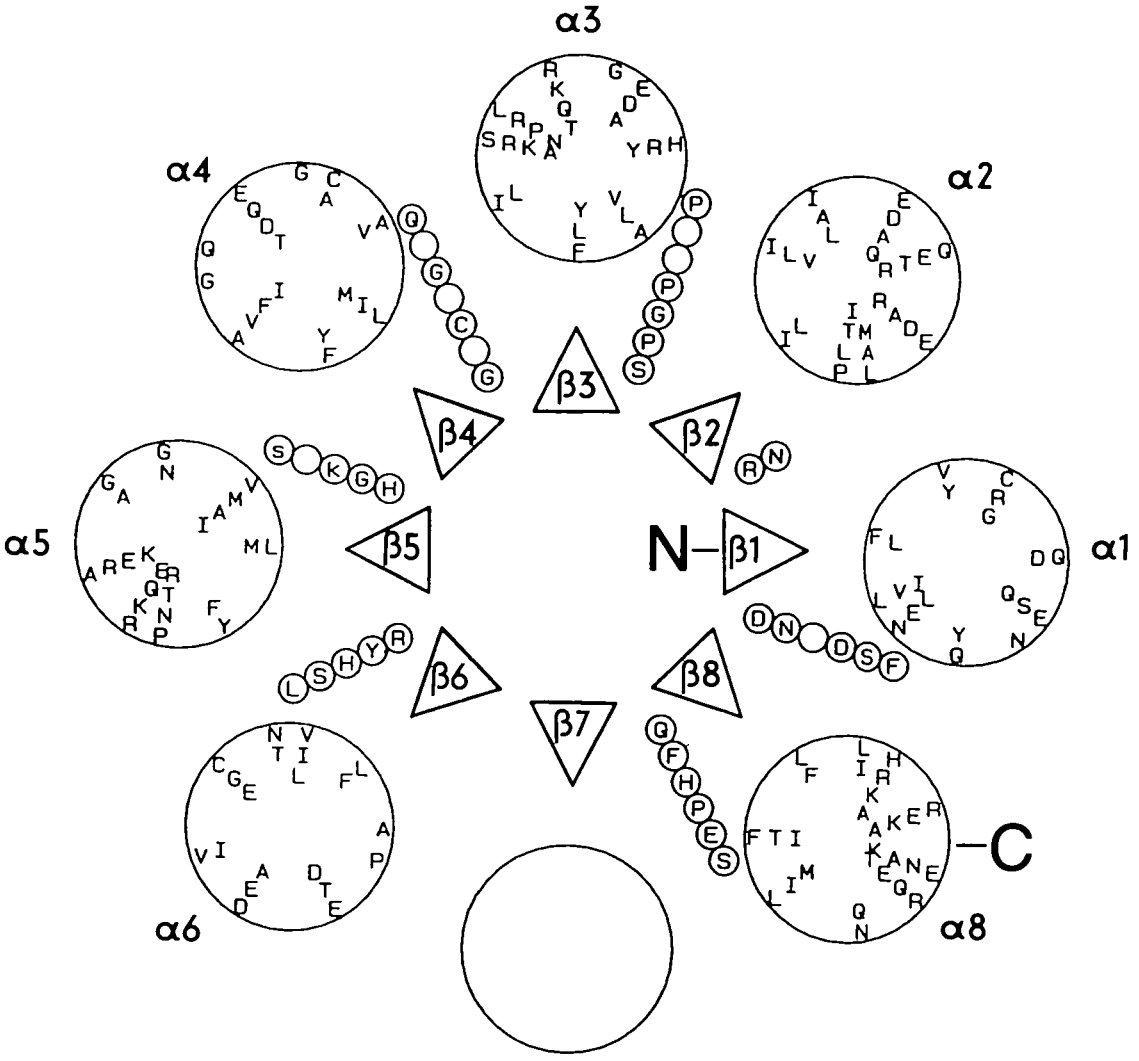


Fig. 2. Arrangement of predicted secondary structural elements and loops into an idealized TIM-barrel. Projection down the central β-barrel axis. N, N-terminus; C, C-terminus. (Δ) Strands β_i, labeled as in Figure 1, C-termini facing upwards. (•) Individual residues of loops connecting C-termini of β-strands with N-termini of α-helices. Single-letter amino acid code indicates conserved residues from Figure 1. (O) Helices α_i labeled as in Figure 1. Single-letter amino acid code indicates predominant residues of the central helix region, presented as a conical helical wheel to illustrate the amphipathic character of the predicted α-helices.

small circles correspond to the invariant residues given in Figure 1. Note that all invariant residues that are candidates for a catalytic function (e.g. general acid/base and nucleophilic catalysis) are located in loops at the C-terminal end of the central β-barrel.

The letters within the large circles correspond to the predominant residues of the predicted α-helices, and are distributed at 100° intervals in a right-handed helical fashion on the surface of a cone descending away from the observer. The seven predicted α-helices decorated in this manner as helical wheels are clearly amphipathic, as generally observed in TIM-barrels. In Figure 2 they are arranged arbitrarily, with their hydrophobic surfaces facing the central β-barrel.

Further support for secondary structural elements that are ambiguously predicted in Figure 1 is derived from the distribution of clusters of conserved residues. The prediction method used here was optimized using multiple sequences from seven different TIM-barrel proteins (Niermann and Kirschner, 1991a). It was found that a template of supersecondary structure occurred frequently along each sequence, and it helped to correct most of the mis- or unpredicted secondary structural elements. The template consists of a hydrophobic β-strand followed by a loop and an amphipathic α-helix. Importantly, conserved residues are frequently clustered at the C-terminus of the β-strand, and in the adjacent loop segment.

Figure 1 shows that the strongly predicted βα units 1, 2, 3, 4 and 6 fit to this template. Moreover, it supports the otherwise ambiguous assignment of β_5. The occurrence of the conserved cluster 104 Gg 105 at the C-terminus of the ambiguously predicted helix α_4 does not necessarily invalidate the assignment, because tight loops between α-helices and β-strands in TIM-barrels frequently belong to the $\alpha\beta_1$ or $\alpha\beta_3$ type, which includes a conserved glycine residue (Scheerlinck *et al.*, 1992). Conserved glycines are also found at the C-termini of α_1 (G25) and α_3 (G85). It is also predicted from the previous analysis of TIM-barrel proteins (Niermann and Kirschner, 1991a) that insertions are not tolerated between the predicted β-strands and the overlapping conserved residues, but rather between these conserved residues and the subsequent α-helices. Figure 1 shows that where aligned gaps do occur between β-strands and α-helices (e.g. $\beta_2\alpha_2$, $\beta_3\alpha_3$, $\beta_5\alpha_5$, $\beta_6\alpha_6$, $\beta_8\alpha_8$), they are always located after the cluster of conserved residues that

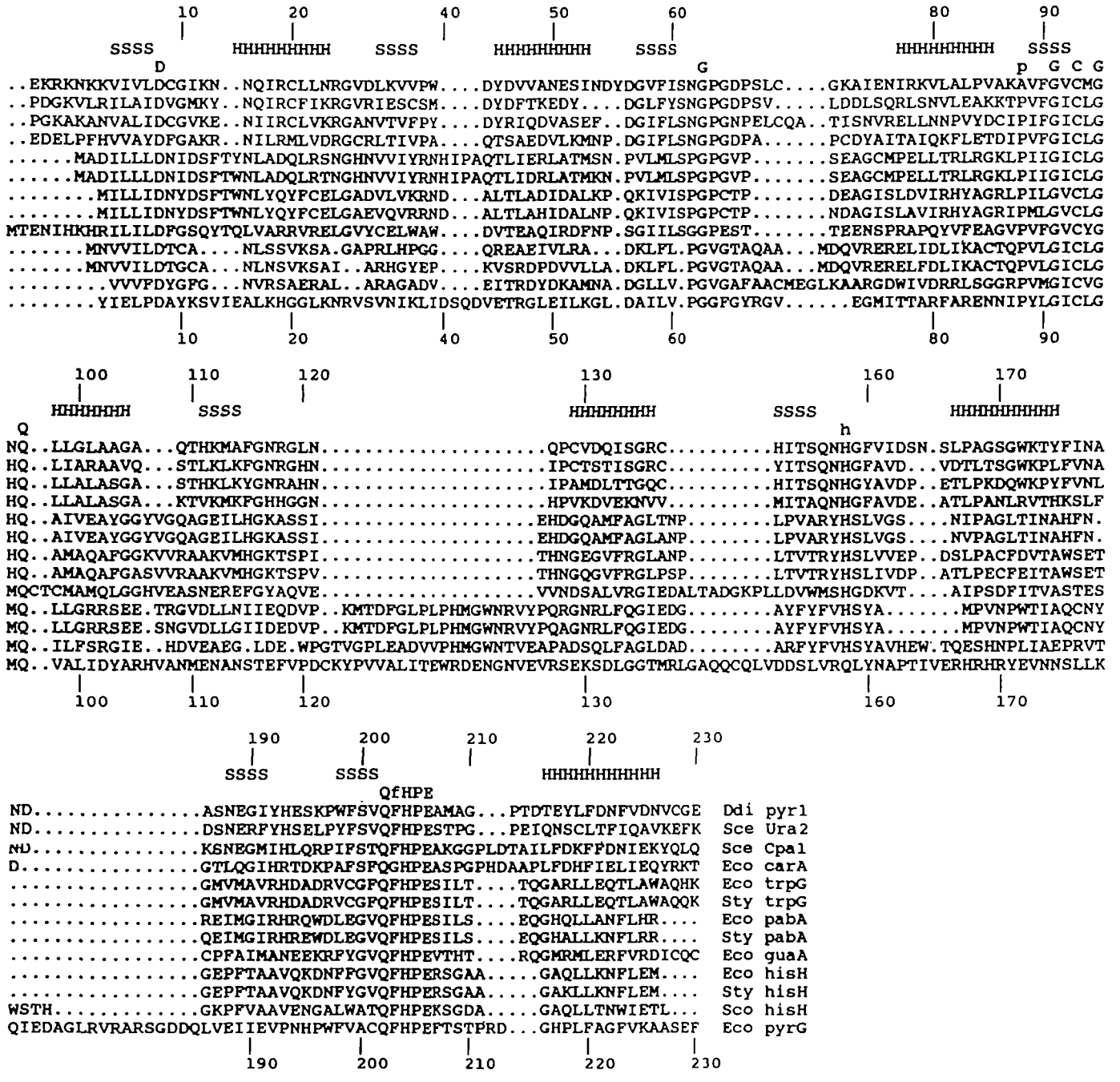


Fig. 3. Illustrative alignment of the superfamily of G-type GATs. Numbering is as in Figure 1. (S) Predicted β-strand position. (H) Predicted α-helix position. The single-letter amino acid code has been used. Top row: conserved residues; upper case, invariant; lower case, allowing for one difference. Section of aligned sequences: *Ddi pyr*I to *Eco car*A, carbamoyl phosphate synthetases; *Eco trp*G and *Sty trp*G, anthranilate synthase components II; *Eco pab*A and *Sty pab*A, aminodeoxychorismate synthases; *Eco gua*A, GMP synthetase; *Eco his*H to *Sco his*H, imidazoleglycerol phosphate synthase; *Eco pyr*G, CTP synthetase. For definitions of acronyms of microorganisms see the text and Materials and methods.

follow the β-strand. Interestingly, the genomic DNA of the fungus *P.chrysosporium* carries an intron at a position corresponding to the peptide bond between A108 and Y109, i.e. between α_4 and β_5 (cf. Figure 1; Schrank *et al.*, 1991).

Helices α_4, α_5 and α_7 are the only underpredicted secondary structural elements expected for a basic TIM-barrel protein. The region between residues 95 and 106 that is a candidate for α_4 is characterized by small amplitudes of the three propensity profiles (Figure 1A), with a preference for β-strand. In contrast, amphipathy and low chain flexibility are the dominant average properties in Figure 1B in this segment. Conversely, the region between residues 107 and 116 that is a candidate for β_5 is predicted ambiguously as either β-strand or α-helix from the propensity profiles in Figure 1A, but again amphipathy and low chain flexibility are the dominant properties. Several members of the TIM-barrel family deviate from the standard fold shown in Figure 2. Thus, the region of residues 95–120 may resemble the unorthodox topology of enolase (Lebioda *et al.*, 1989) in having the secondary structure $\beta_4\beta_5\alpha_4\alpha_5$, rather than $\beta_4\alpha_4\beta_5\alpha_5$ as given in Figure 1. Similarly, the missing α-helices in the TIM-barrels of mandelate racemase and muconate lactonizing enzyme (Neidhart *et al.*, 1990),

phosphoribosyl anthranilate isomerase (Wilmanns *et al.*, 1992) and *N*-acetylglucosaminidase (Van Roey *et al.*, 1994) are precedents for the underpredicted helix α_7 of the GAT domain.

Information from the G-type GAT superfamily

Chemical labeling and mutational studies have provided direct evidence that the invariant residues cysteine C93, histidine H204 and glutamate E206 (numbering as in Figure 1) are catalytically essential for the G-type GAT domains of both anthranilate (*trp*G) and aminodeoxy chorismate (*pab*A) synthases (Roux and Walsh, 1992, 1993; Zalkin, 1993). It has been suggested that these three residues cooperate as a 'catalytic triad' during the hydrolysis of the carboxamide bond of glutamine. In any case, all three residues would be located at the C-terminal face of the proposed TIM-barrel of the G-type GAT domain (Figure 2).

To gain further insight into the subset of conserved residues that affect the intrinsic activity of the G-type GAT domain, we compared selected sequences of the corresponding superfamily (Zalkin *et al.*, 1985; Zalkin, 1993) with the sequences and predicted secondary structural elements of the *trp*G/*pab*A subfamily given in Figure 1. For the sake of comparison, the arbitrary numbering of Figure 1 is also used in Figure 3. The alignment contains four sequences of carbamyl phosphate synthetase (*Ddi pyr*1 from *Dictyostelium discoideum*; Faure *et al.*, 1989; URA2, CPA1 and *car*A; Zalkin, 1993), two representative sequences each of anthranilate synthase (*trp*G) and aminodeoxychorismate (*pab*A) synthases from Figure 1, a single sequence of GMP synthase (*gua*A from *E.coli*; Mäntsälä and Zalkin, 1992) out of four, three sequences of the GAT domain of imidazolglycerol phosphate synthase (*his*H; Kuenzler *et al.*, 1993), and a single sequence of CTP synthetase (*pyr*G from *E.coli*; Zimmer and Hundeshagen, 1994) out of four. The two known sequences of the GAT domain of 5′-phosphoribosylformyl-glycinamide amidotransferase [product of the bifunctional *pur*L(Q) gene of *E.coli* and the monofunctional *pur*Q gene of *B.subtilis*] showed significant sequence identities in only two regions (residues 91–97 and 202–207 in Figure 3; Sampei and Mizobuchi, 1989; Schendel *et al.*, 1989), and were therefore not included in the alignment.

On one hand, several of the aligned gaps in Figure 1 have disappeared (e.g. around position 30), because only four out of the 26 sequences of *trp*G and *pab*A from Figure 1 were used for the comparison. On the other hand, the alignment in Figure 3 required the introduction of five new aligned gaps around positions 15, 55, 61, 97 and 105. Although the alignments are equivocal in certain regions, it is reassuring that all the newly introduced gaps lie between and not within predicted secondary structural elements.

Five out of seven groups of invariant residues in Figure 1 are retained in Figure 3: (I) D8; (III) G63, (IV) 91 GxCxGxQ 97; (VII) H158 and (VIII) 202 QxHPE 206. Based on the joint prediction of Figure 1, all five groups are located a few residues beyond the C-termini of the predicted strands β_1, β_3, β_4, β_6 and β_8. Weng and Zalkin (1987) have studied the role of G63 in the G-type GAT domain of CTP synthetase from *E.coli* (*pyr*G). The corresponding mutant G351A appeared to be unstable. The histidine residue in the GAT domain of carbamoyl phosphate synthetase from *E.coli* (*car*A) that corresponds to H158 was mutated by Miran *et al.* (1991) to asparagine (H312N). Because k_{cat} was unaffected but the K_M value for glutamine was increased 100-fold, it was suggested that H312 is involved in the binding of the substrate. Similarly, Mareya and Raushel (1994) showed by mutagenesis of C248 in the same enzyme that bulky replacements increase the unproductive glutaminase activity but abolish the transfer of the glutamine amino group to carbamyl phosphate. C248 corresponds to the variable position 72 in Figures 1 and 3, which is located in the loop between β_3 and α_3. Clearly this loop must be involved both in the intrinsic hydrolytic activity of the GAT domain and in its interaction with the synthase domain of carbamoyl phosphate synthase.

To our knowledge, the invariant residues D8, G91, G94, Q96, Q202 and P205 in the G-type GAT domain have not been challenged as yet by mutagenesis. Because E170 in the GAT domain of aminodeoxychorismate synthase (which corresponds to E206 in Figures 1 and 3) can be replaced by aspartate with the retention of 25% catalytic efficiency (Roux and Walsh, 1993), it remains to be seen whether D8 is an alternative general base in the catalytic triad. If the G-type GAT domain possesses the TIM-barrel fold, strand β_1 followed by D8 would be in close proximity to strand β_8 followed by 602 QxHPE 206 (see Figure 2). In summary, several residues that affect indirectly the function of the various GAT domains of the superfamily are also located in $\beta_i\alpha_i$ loops of the postulated TIM-barrel fold.

The α/β hydrolase fold described recently comprises eight β-strands and six α-helices (Ollis *et al.*, 1992). This superfamily of enzymes utilizes a catalytic triad of Asp/Glu–His–Ser/Cys to catalyze the hydrolysis of a great variety of compounds, including peptides and esters. Moreover, the C-terminal half of the fold comprises four parallel $\beta\alpha$ modules that carry the residues of the catalytic triad. Nevertheless, the α/β hydrolase fold is not a reasonable alternative topology for GAT because the four N-terminal $\beta\alpha$ motifs of GAT are strongly predicted (see Figure 1), whereas the α/β hydrolase fold has the four consecutive $\beta\alpha$ motifs at its C-terminus [secondary sequence $\beta\beta(\beta\alpha)_6$]. The catalytic triad of the α/β hydrolase fold is located at the ends of the following β-strands: β_5 (Cys/Ser), β_7 (Asp/Glu) and β_8 (His).

Results from alternative prediction methods

The earlier mean secondary structure prediction by Zalkin *et al.* (1985), which was based on the alignment of single sequences each of anthranilate synthase (*trp*G), aminodeoxychorismate synthase (*pab*A) and GMP synthetase (*gua*A), differs from our joint prediction in $\beta\alpha$ units 1, 3, 5, 6, 7 and 8, but does predict α_7 in the segment 193–198. Wilmanns and Eisenberg (1993) used the 3-D profile method on a large database of protein sequences of unknown structure. The GAT sequence was found to be compatible with the profiles of the TIM-barrel proteins triose phosphate isomerase and indoleglycerol phosphate synthase. Pickett *et al.* (1992) tested the sequence template method for discriminating the TIM-barrel fold on a limited database that unfortunately did not contain a GAT sequence, perhaps because some sequences are <200 residues in length (Figure 1).

Conclusion

The predicted secondary structural elements of the G-type GAT domains of the anthranilate (*trp*G) and aminodeoxychorismate (*pab*A) synthases fall into the pattern of eight consecutive β-strand–loop–α-helix–loop modules that are typical for TIM-barrel topology. This working model is supported further by the location of (i) aligned gaps and clusters of invariant residues and (ii) the known catalytically essential residues Cys93, His204 and Glu206 in loops at the C-terminal face of the 8-fold parallel β-barrel. Because the sequences of four

members of the GAT domain superfamily can be aligned to the predicted secondary structural elements, several additional residues involved indirectly in substrate binding and the transfer of 'nascent NH_3' are predicted to reside in the neighborhood of the catalytic triad. The forthcoming elucidation of the structure of GMP synthase, which contains a G-type GAT domain (Tesmer *et al.*, 1994), will decide the issue.

Acknowledgements

We thank E.Johner for assembling the manuscript. The work was supported by the Swiss National Science Foundation grant no. 31-25711.88.

References

Benner,S.A. (1992) *Curr. Opin. Struct. Biol.*, **2**, 402–412.
Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) *J. Mol. Biol.*, **195**, 659–685.
Crawford,I.P. (1989) *Annu. Rev. Microbiol.*, **43**, 567–600.
Crawford,I.P., Niermann,T. and Kirschner,K. (1987) *Proteins*, **2**, 118–129.
Eisenberg,D., Weiss,R.M. and Terwilliger,T.C. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 140–144.
Essar,D.W., Eberly,L., Han,D.-Y. and Crawford,I.P. (1990) *J. Bacteriol.*, **172**, 853–866.
Farber,G.K. (1993) *Curr. Opin. Struct. Biol.*, **3**, 409–412.
Faure,M., Camonis,J.H. and Jacquet,M. (1989) *Eur. J. Biochem.*, **179**, 345–358.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.
GCG (1991) *Program Manual for the Wisconsin Package. Version 7 April 1991.* Genetics Computer Group, 575 Science Drive, Madison, WI 53711, USA.
Gibrat,J.-F., Garnier,J. and Robson,B. (1987) *J. Mol. Biol.*, **198**, 425–443.
Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
Kaplan,J.B., Merkel,W.K. and Nichols,B.P. (1985) *J. Mol. Biol.*, **183**, 327–340.
Karplus,P.A. and Schulz,G.E. (1985) *Naturwissenschaften*, **72**, 212–213.
Kuenzler,M., Balmelli,T., Egli,C.M., Paravicini,G. and Braus,G. (1993) *J. Bacteriol.*, **175**, 5548–5558.
Kyte,J. and Doolittle,R.F. (1982) *J. Mol. Biol.*, **157**, 105–132.
Lasters,I., Wodak,S.J. and Pio,F. (1990) *Proteins: Struct. Funct. Genet.*, **7**, 249–256.
Lebioda,L., Stec,B. and Brewer,J.M. (1989) *J. Biol. Chem.*, **264**, 3685–3693.
Lesk,A.M., Brändén,C.-I. and Chothia,C. (1989) *Proteins: Struct. Funct. Genet.*, **5**, 139–148.
Levitt,M. and Chothia,C. (1976) *Nature*, **261**, 552–557.
Mäntsälä,P. and Zalkin,H. (1992) *J. Bacteriol.*, **174**, 1883–1890.
Mareya,S.M. and Raushel,F.M. (1994) *Biochemistry*, **33**, 2945–2950.
Mewes,H.W. (1990) Max-Planck-Institute protein sequences. Max-Planck-Institute für Biochemie, Munich, Germany.
Miran,S.G., Chang,S.H. and Raushel,F.M. (1991) *Biochemistry*, **30**, 7901–7907.
Murzin,A.G., Lesk,A.M. and Chothia,C. (1994a) *J. Mol. Biol.*, **236**, 1369–1381.
Murzin,A.G., Lesk,A.M. and Chothia,C. (1994b) *J. Mol. Biol.*, **236**, 1382–1400.
Neidhart,D.J., Kenyon,G.L., Gerlt,J.A. and Petsko,G.A. (1990) *Nature*, **347**, 692–694.
Niermann,T. and Kirschner,K. (1991a) *Protein Engng*, **4**, 359–370.
Niermann,T. and Kirschner,K. (1991b) *Methods Enzymol.*, **202**, 45–59.
Ollis,D.L. *et al.* (1992) *Protein Engng*, **5**, 197–211.
Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Pickett,S.D., Saqi,M.A.S. and Sternberg,M.J.E. (1992) *J. Mol. Biol.*, **228**, 170–187.
Roux,B. and Walsh,C.T. (1992) *Biochemistry*, **31**, 6904–6910.
Roux,B. and Walsh,C.T. (1993) *Biochemistry*, **32**, 3763–3768.
Sampei,G. and Mizobuchi,K. (1989) *J. Biol. Chem.*, **264**, 21230–21238.
Scheerlinck,J.-P.Y., Lasters,I., Claessens,M., De Mayer,M., Pio,F., Delhaise,P. and Wodak,S.J. (1992) *Proteins: Struct. Funct. Genet.*, **12**, 299–313.
Schendel,F.J., Mueller,E., Stubbe,J., Shiau,A. and Smith,J.M. (1989) *Biochemistry*, **28**, 2459–2471.
Schrank,A., Tempelaars,C., Sims,P.F.G., Oliver,S.G. and Broda,P. (1991) *Mol. Microbiol.*, **5**, 467–475.
Smith,J.L., Zaluzec,E.J., Wery,J.-P., Niu,L., Switzer,R.L., Zalkin,H. and Satow,Y. (1994) *Science*, **264**, 1427–1433.
Tesmer,J.J.G., Stemmler,T.L., Penner-Hahn,J.E., Davisson,V.J. and Smith,J.L. (1994) *Proteins: Struct. Funct. Genet.*, **18**, 394–403.
Tso,J. and Zalkin,H. (1980) *J. Biol. Chem.*, **255**, 1451–1457.
Van Roey,P., Rao,V., Plummer,T.H.,Jr and Tarentino,A.L. (1994) *Biochemistry*, **33**, 13989–13996.
Weng,M. and Zalkin,H. (1987) *J. Bacteriol*, **169**, 3023–3028.
Wilmanns,M. and Eisenberg,D. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 1379–1383.
Wilmanns,M., Priestle,J.P., Niermann,T. and Jansonius,J.N. (1992) *J. Mol. Biol.*, **223**, 477–507.
Wu,T.T. and Kabat,E.A. (1970) *J. Exp. Med.*, **132**, 211.
Zalkin,H. (1993) *Adv. Enzymol. Relat. Areas Mol. Biol.*, **66**, 203–310.
Zalkin,H., Argos,P., Naranyana,S.V.L., Tiedeman,A.A. and Smith,J.M. (1985) *J. Biol. Chem.*, **260**, 3350–3354.
Zimmer,W. and Hundeshagen,B. (1994) *FEMS Microbiol. Lett.*, **115**, 273–278.
Zvelebil,M.J., Barton,G.J., Taylor,W.R. and Sternberg,M.J.E. (1987) *J. Mol. Biol.*, **195**, 957–961.

Received January 4, 1995; revised March 27, 1995; accepted April 26, 1995

Note added in proof

The G-type glutamine amidotransferase domain of GMP synthase does not have the TIM barrel topology. The predicted assignments of the first four β-strands and α-helices (positions 1–104 in Figure 2) and of the last two β-strands and the last α-helix (positions 188–227) are essentially correct. However, the assignments of the intervening region (positions 105–187) are largely incorrect (Dr Janet Smith, personal communication).